Northeastern University, Boston, MA

College of Engineering

Department of Civil and Environmental Engineering

CIVE 7381: Transportation Demand Forecasting and Model Estimation

Problem Set 3

From: **Nathan David Obeng-Amoako** (NUID: 002607282)

To: Professor Haris Koutsopoulos

Submitted on: Friday, October 4th

Fall 2024

Table of Contents

Northeastern University
Civil and Environmental
Engineering

List of Figures

List of Tables

List of Equations

List of Abbreviations

OLS = Ordinary Least Squares

RMSE = Root Mean Squared Error

Problems

Northeastern

Department of Civil and
Environmental Engineering
400 Snell Engineering
360 Huntington Ave.
Boston, MA 02115

CIVE 7381 Transportation Demand
Problem Set #3
Due: Wednesday, October 16, 2024

Problem 1

Consider a model that predicts the accident rates for different states in the U.S.:

$$Y_n = \alpha + \beta X_n + \varepsilon_n, \qquad n = 1,\ldots,50$$

where n are indices of states, X_n is the proportion of automobiles exceeding 55 miles per hour in state n, and Y_n is the number of fatalities per million vehicle miles.

The sample means of X and Y are 0.6 and 1.0 respectively. The sum of the squared deviations of X from its mean is 10.0, the sum of the squared deviations of Y from its mean is 1.0, and the sum of cross products of deviations of X and Y from their respective means is 2.0. In other words:

$$\sum_n (X_n - \overline{X})^2 = 10.0$$

$$\sum_n (Y_n - \overline{Y})^2 = 1.0$$

$$\sum_n (X_n - \overline{X})(Y_n - \overline{Y}) = 2.0$$

a) Compute the least squares estimates of α and β, the sum of squared errors (SSE), and R^2.

b) Test the hypothesis H_0: $\beta = 0$ against the H_1: $\beta \neq 0$.

Problem 2

Trip generation models for home-based social recreation purposes have been developed for a metropolitan area. Models have been developed separately for a household based on the number of workers in the household (WHH) and number of cars in the household (AO). The models for households with WHH = 1 and AO = 1 and households with WHH = 2+ and AO = 1 are given below along with the t-statistics (in parentheses) of the various parameters and their R^2 values.

	Model 1	**Model 2**
Parameter	**WHH = 1 and AO = 1**	**WHH = 2+ and AO = 1**
Constant (intercept)	1.014 (4.23)	
HHSIZE		-0.0625 (-1.99)
Log(HHSIZE)	0.3638 (0.98)	
Log(HHInc)		0.23 (6.78)
AreaType	-0.651 (-1.56)	
AreaType^2	0.252 (0.23)	
AreaType^3	-0.028 (-1.35)	
HBWTDUR		-0.00427 (-4.58)
Log(HBWTDUR)	-0.086 (-3.58)	
Sample Size	1720	500
R^2	0.0458	0.2021

Values in parentheses represent the t-statistics from the regression analysis.

HHSize = Household Size
LogHHSize = Natural Log of Household Size
LogHHInc = Natural Logarithm of Household Income in 1000s of $ (MAX(0, LN(Income))
AreaType = Density Based Area Type (see below for definition)
AreaType^2 = Density Based Area Type, Squared
AreaType^3 = Density Based Area Type, Cubed
HBWTDUR = Average One-Way Home-based work (HBW) trip duration, in minutes
LogHBWTDUR = Natural Log of Average One-Way HBW Trip Duration (max(0, *ln*(HBWTDUR))

To determine the area type, the area density is first calculated as:

*Area Density = (Total Population + 2.5 * Total Employment / Developed Acres)*

Then, based on the value of *Area Density,* the Area Type of a zone is determined according to the following table:

	AREATYPE	*Area Density*
0	*Regional Core*	*> 300.0*
1	*CBD*	*100.0 - 300.0*
2	*Urban Business*	*55.0 - 100.0*
3	*Urban*	*30.0 - 55.0*
4	*Suburban*	*6.0 - 30.0*
5	*Rural*	*< 6.0*

The above models were calibrated using data from individual household surveys. Discuss and evaluate the two models in detail:

a) What are the explanatory variables and what do they intend to capture?
b) Is the impact of the explanatory variables reasonable (based on the estimation results)?
c) Comment on the t-statistics of the variables and their interpretation
d) How else you would include the variable *AREATYPE* in the model?
e) Comment on the R^2 values. Calculate the corrected R^2 for both models. How do they compare to the corresponding R^2?
f) What other variables would you include in the model (assuming the data is available)?
g) Does it make sense that different explanatory variables were used to capture the number of trips for the two different household types? Explain why yes, or why not.
h) Would you use these models for a planning study?

Problem 3

This problem is based on problem 2 from PS 2.
You are given a set of aggregate data from 57 traffic analysis zones (TAZ) in the Chicago Area (the data is quite old as you can infer from the "official" description of the variables included in the table in the next page). For each of the 57 zones you have available the average trips per occupied dwelling unit, the average car ownership, the average household size, and three zonal social indices. The data is the same data you used in PS 2.

Based on your preliminary analysis on PS 2, develop linear regression models that predict the auto trips/day generated by a household. Present the best two model specifications you came up with and interpret the parameters and their signs. Report all relevant results from statistical tests.

You may use any statistical software (including R and python) to estimate the models (you can also use Excel using the regression analysis function).

The data file contains the following variables that can be used for the specification of your model:

Name	Description
TODU	*Trips per Occupied Dwelling Unit* Trips refer to the daily frequency of person-trips via motor vehicle (auto driver or passenger) or public transit made from a dwelling unit by members of that dwelling unit. All trips whose origins were other than "from home" were ignored.
ACO	*Average Car Ownership* Cars per dwelling unit.
AHS	*Average Household Size* Number of residents per dwelling unit.
SRI	*Job/Skills Rank Index* This index reflects two elements: (i) the proportion of blue-collar workers, defined as the ratio of craftsmen, operatives, and laborers to all employees; and (ii) educational level as measured by the proportion of persons 25 years and older completing eight or fewer years of schooling. The index attains a maximum value when no residents are in the blue-collar jobs category, and all adult residents have more than eight years of education
UI	*Urbanization Index* This index reflects three elements: (i) the ratio of children under five years of age to the female population of childbearing age; (ii) percentage of women who are in the labor force, and (iii) the percentage of single units to total dwelling units. The degree of urbanization index would be increased by (a) lower ratio in (i), (b) higher percentage in ii); and (c) lower proportion of single dwelling units. High values for this index imply less attachment to the home because of fewer children, higher likelihood of women being employed, and less permanency of dwelling unit type in terms of average tenure.
MI	*Minority Index* This index is defined as the proportion of an area's residents who are minorities.

Solution to Problem 1

Problem 1a (Coefficients, SSE and R^2)

Coefficients

The OLS estimates of the slope (β) and intercept (α) calculated using **Equation 1** and **Equation 2** are **0.20** and **0.88**, respectively.

$$\beta = \frac{Cov(X,Y)}{Var(X)} = \frac{\sum_n (X_n - \bar{X})(Y_n - \bar{Y})}{\sum_n (X_n - \bar{X})^2}$$

Equation 1: Slope of a simple linear regression model

Where:

$Cov(X,Y)$ = covariance between X and Y

$Var(X)$ = variance of X

$$\alpha = \bar{Y} - \beta\bar{X}$$

Equation 2: Intercept of a simple linear regression model

Where:

$\bar{X}$ = sample mean of X

$\bar{Y}$ = sample mean of Y

Hence, the calibrated linear regression model is $Y_n = 0.88 + 0.20X_n$

Sum of squared errors (SSE)

From **Equation *3***, the SSE is **0.60**.

$$SSE = \sum_n (Y_n - \bar{Y})^2 - \beta \sum_n (X_n - \bar{X})(Y_n - \bar{Y})$$

Equation 3: Sum of squared errors

Coefficient of Determination (R^2)

From **Equation *4***, the coefficient of determination, R^2, is **0.40**.

$$R^2 = 1 - \frac{SSE}{\sum_n (Y_n - \bar{Y})^2}$$

Equation 4: Coefficient of Determination

Problem 1b (Hypothesis Testing)

Null hypothesis, $H_0: \beta = 0$

Alternate hypothesis, $H_1: \beta \neq 0$

T-statistic for $\boldsymbol{\beta}$

Assuming the error is normally distributed, the t-statistic for β from **Equation *5*** is **5.6569**.

$$t = \frac{b-\beta}{s_\beta} = \frac{b-\beta}{s_\varepsilon / \sqrt{\sum_n (X_n - \bar{X})^2}}$$

Equation 5: T-statistics for the slope, β

Where:

$b - \beta$	= difference between estimated and hypothesized slopes
s_β	= standard error of β
$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$	= residual standard error ($n-2$ is the degree of freedom)

I choose to test my hypotheses at the 95% confidence level. At this level and with 48 degrees of freedom, the critical value from the t-distribution is approximately 2. Since t > 2, we reject the null hypothesis. The interpretation of this result is that the proportion of automobiles exceeding 55 mph (X_n) has a **statistically significant** effect on the number of fatalities per million vehicle miles (Y_n) across the 50 states in America.

Solution to Problem 2

The functions in **Figure 1** were plotted in Python to kind of visualize the explanatory variables and help answer the questions.

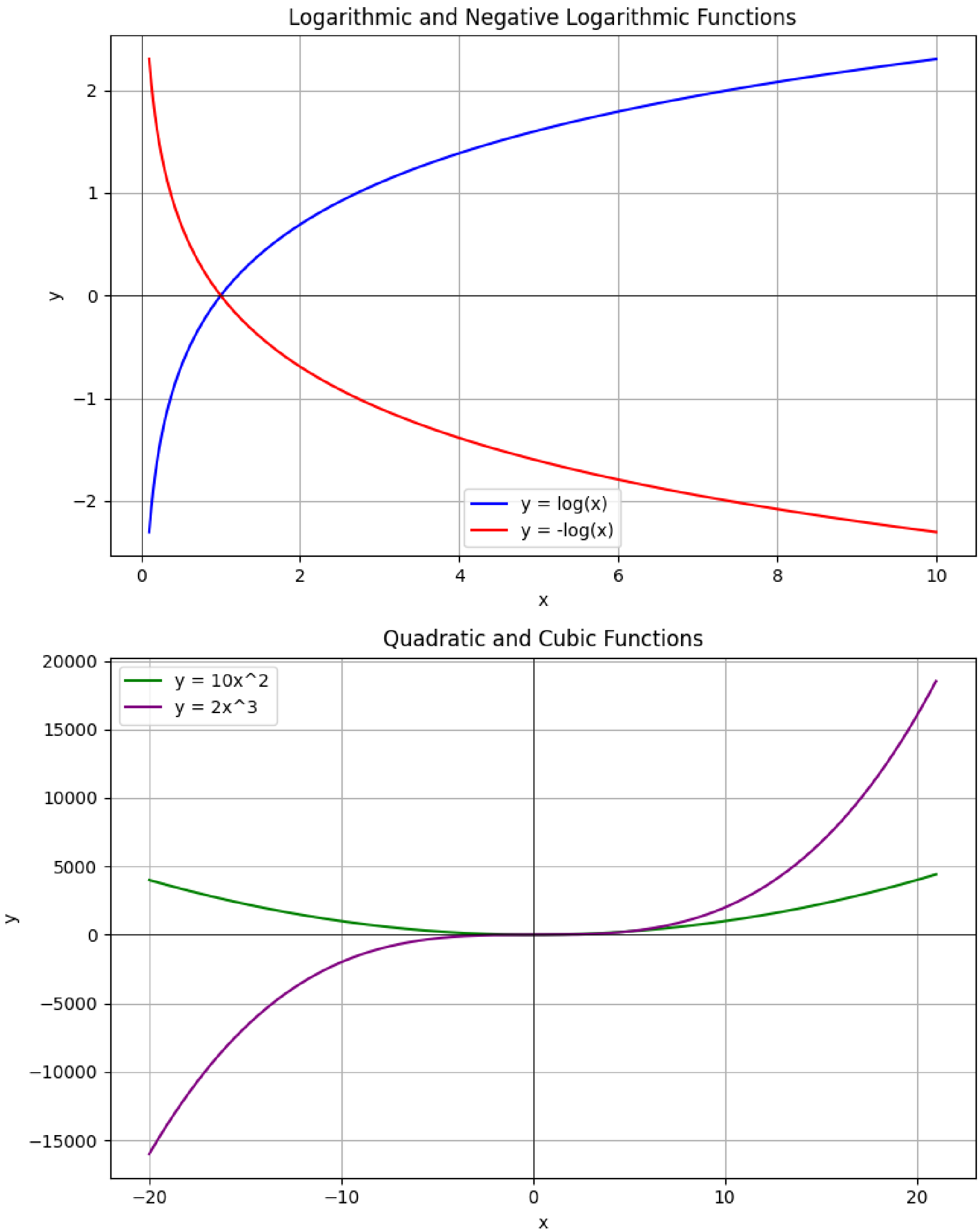

Figure 1: Log, Neg Log, Quadratic and Cubic Functions

Problem 2a (Explanatory Variables)

Model 1 (WWH = 1 and AO = 1)

- **log(HHSIZE):** The positive coefficient of the logarithmic transformation of HHSIZE captures increase in number of home-based social recreation trips at a diminishing rate as HHSIZE increases.

- **AreaType:** The negative coefficient of AreaType capture a linear relationship which is such that households in denser areas generate more home-based social recreation trips.

- **AreaType^2** and **AreaType^3:** Both of these explanatory variables try to capture non-linear relationship between area type and number of home-based social recreation trips generated.

- **log(HBWTDUR):** Similar to log(HHSIZE), this explanatory variable also captures how average one-way HBW trip durations affect trip generation. Because of the negative sign, as HBWTDUR increases, trip generation decreases at a diminishing rate.

Model 2 (WWH = 2+ and AO = 1)

- **HHSIZE:** The negative coefficient suggests that larger households might make fewer social recreation trips.

- **log(HHInc)**: The positive coefficient of the logarithmic transformation of HHInc suggests that social trips increase at a diminishing rate as household income increases.

- **HBWTDUR:** The negative coefficient of HBWTDUR capture a linear relationship which is such that longer home-based work trip durations reduce the likelihood of social trips.

Problem 2b (Impact of Explanatory Variables)

The only variables of the ones listed in part a) that are unreasonable are **HHSIZE** for Model 2 and **AreaType^2** for Model 1. General expectations have been listed below:

- Increase in **log(HHSIZE), HHSIZE,** and **log(HHInc)** are expected to result in some form of increase in home-based social recreation trips.

- Increase in **AreaType, AreaType^2**, **AreaType^3, HBWTDUR,** and **log(HBWTDUR)** are expected to result in some form of decrease in home-based social recreation trips.

Problem 2c (t-statistics)

The null hypotheses being tested are whether the estimated coefficient of the respective explanatory variable is equal to zero. At the 95% confidence level, a coefficient with a magnitude greater than 2 means that we reject the null hypothesis that the estimated coefficient is equal is zero. What this implies is that the explanatory variable has strong statistical significance and can be used to predict trip generation. On the other hand, if the magnitude of the t-statistic is less than 2, then we fail to reject the null hypothesis, which implies that the variable might be irrelevant when it comes to predicting the number of home-based social recreation trips.

Model 1 (WWH = 1 and AO = 1)

- **Intercept (4.23):** This t-statistic is statistically significant. Gives a good idea of the baseline trip generation with all explanatory variables equal zero.

- **log(HHSIZE) (0.98):** This t-statistic of 0.98 is not statistically significant at the 95% confidence level. This means that log(HHSIZE) is not a reliable predictor variable.

- **AreaType (- 1.56):** Does not reach significance at the 95% confidence level.

- **AreaType^2 (0.23):** Very low t-statistic implies that **AreaType^2** is not a good explanatory variable to predict number of home-based social recreation trips.

- **AreaType^3 (-1.35):** Low t-statistic implies AreaType^3 is not significant.

- **log(HBWTDUR) (-4.58):** T-statistic is much higher than 2 which implies that the explanatory variable is statistically significant. This suggests that the logarithmic transformation of the average one-way HBW trip durations is very relevant in predicting number of home-based social recreation trips.

Model 2 (WWH = 2+ and AO = 1)

- **HHSIZE (-1.99):** This t-statistic is only slightly less than 2. This means that at the 95% confidence level, household size might be able to predict trip generation but it is only marginally significant.

- **log(HHInc) (6.78):** The t-statistic is much greater than 2, which indicates very strong significance of the explanatory variable, log(HHInc), in predicting the response variable, number of home-based social recreation trips.

- **HBWTDUR (-4.58):** T-statistic is much higher than 2. This suggests that average one-way HBW trip durations is very relevant in predicting number of social recreation trips.

Problem 2d (AREATYPE Transformations)

Area type is a categorical variable. It'll only make sense to represent it with multi-level dummy variable(s) rather than trying to treat it like a numerical variable.

Problem 2e (Corrected R^2 Values)

From **Equation 6**, the corrected R^2 for **Model 1** and **Model 2** are 0.0436 and 0.19989, respectively.

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$$

Equation 6: Corrected R^2

Where:

R^2 = coefficient of determination
n = number of data points
k = number of independent variables

Problem 2f (Additional Explanatory Variables)

Age distribution, employment status, access to public transportation, land use mix, and type of occupation could all be good explanatory variables for the trip generation models.

Problem 2g (Why Different Explanatory Variables?)

Trip generation can vary significantly across household types, making it easier to predict for some than others. As a result, different household types with distinct travel behaviors may require different number of explanatory variables reflecting the complexity or simplicity of their travel patterns. Additionally, certain variables that are critical for one household type may be irrelevant for another, justifying the need for specific sets of explanatory variables.

Problem 2h (Planning Study)

I personally will not use these models for a planning study partly based on the low coefficient of determination (R^2) values which measure the proportion of variation in social trip generation that is explained by the explanatory variables in the regression models. The R^2 values for Model 1 and Model 2 are less than 5% and less than 20%, respectively. Model 2 could be better but both are weak and unreliable, and may not give accurate predictions.

Solution to Problem 3 (Linear Regression Models)

In all, **five** models with different combinations of explanatory variables were tested. The models have been listed below. Note that **Model 4** includes the interaction term between **ACO and UI** to try to capture non-linear effects. Also, **Model 5** is the **full** multiple linear regression developed with all the variables. These models were chosen to be tested based on the scatter and correlation matrices between the dependent and independent variables.

Models Tested

Model 1: $TODU = \beta_0 + \beta_1 \times ACO + \varepsilon$

Model 2: $TODU = \beta_0 + \beta_1 \times UI + \varepsilon$

Model 3: $TODU = \beta_0 + \beta_1 \times ACO + \beta_2 \times UI + \varepsilon$

Model 4: $TODU = \beta_0 + \beta_1 \times ACO + \beta_2 \times UI + \beta_3 \times ACO \times UI + \varepsilon$

Model 5: $TODU = \beta_0 + \beta_1 \times ACO + \beta_2 \times AHS + \beta_3 \times MI + \beta_4 \times SRI + \beta_5 \times UI + \varepsilon$

Model Evaluation Results

Table 1 shows a summary of the OLS regression results for all the five models. The full results are shown in the **Appendix**.

Table 1: Summarized Model Evaluation Results

	Model 1	Model 2	Model 3	Model 4	Model 5
R^2	0.617	0.573	0.695	0.698	0.704
Corrected R^2	0.610	0.565	0.684	0.681	0.675

Figure 2 illustrates the relationship between predicted and actual values for the five regression models.

Discussion of Results

The results of **Model 1** and **Model 2** show that both **ACO** and **UI**, when considered separately, are strong and statistically significant predictors of **TODU**, with a positive and negative relationship, respectively. **Model 1** explains **61.0%** while **Model 2** explains **56.5%** of the variance in **TODU**.

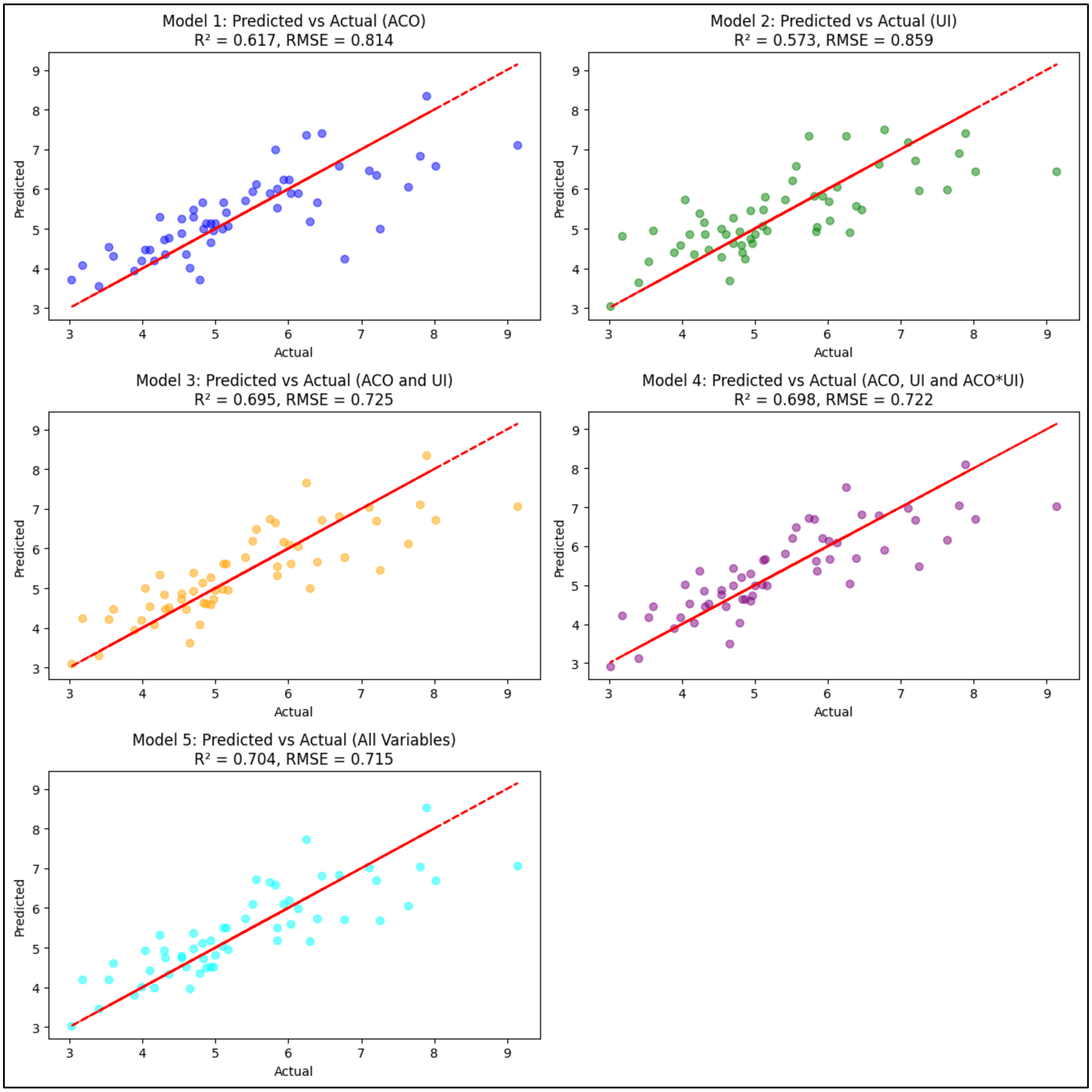

Figure 2: Scatter Plot for Predicted vs Actual

Model 3 with both **ACO** and **UI** as independent variables fits even better. In fact, it's the best of the five models I tested. This model can explain **68.4%** of the variability in **TODU**. The RMSE is also lower. To add to that, the t-statistics used to test the null hypotheses that each of the coefficients of this model is equal to zero (i.e., the respective variable has no effect on the dependent variable) fell outside the rejection region, which means that both independent variables are statistically significant in the model.

Model 4 included an interaction term (**ACO×UI**). The t-statistic of the coefficient of this term suggests that the interaction term is not statistically significant in the model. Though it is the second best-performing (corrected R^2 of 68.1%) of the five models I chose, it doesn't perform any better than the same model but without the interaction term, which makes sense.

Model 5 is very complex (or it is not parsimonious, as you'll put it). I wasn't expecting it to even produce excellent results in the first place. I'll expect this model to lack generalizability. Even though the corrected R^2 value (**67.5%**) is not too bad, I'd guess that it's only performing well on the training data and is likely to perform poorly on unseen data due to overfitting.

Based on the discussion above, **Model 3** and **Model 4** have emerged as the best two models (with the very best being **Model 3**). In both of these two models, the signs of the coefficients of **ACO** and **UI** are what I'd expect.

Appendix

MODEL 1 – FULL REGRESSION RESULTS

```
Model 1: Using ACO
                            OLS Regression Results
==============================================================================
Dep. Variable:                   TODU   R-squared:                       0.617
Model:                            OLS   Adj. R-squared:                  0.610
Method:                 Least Squares   F-statistic:                     88.43
Date:                Fri, 04 Oct 2024   Prob (F-statistic):           4.82e-13
Time:                        18:38:09   Log-Likelihood:                -69.118
No. Observations:                  57   AIC:                             142.2
Df Residuals:                      55   BIC:                             146.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6177      0.517      1.194      0.238      -0.419       1.655
ACO            5.8582      0.623      9.404      0.000       4.610       7.107
==============================================================================
Omnibus:                       15.692   Durbin-Watson:                   1.616
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.800
Skew:                           1.213   Prob(JB):                     0.000136
Kurtosis:                       4.270   Cond. No.                         9.49
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

MODEL 2 – FULL REGRESSION RESULTS

```
Model 2: Using UI
                            OLS Regression Results
==============================================================================
Dep. Variable:                   TODU   R-squared:                       0.573
Model:                            OLS   Adj. R-squared:                  0.565
Method:                 Least Squares   F-statistic:                     73.80
Date:                Fri, 04 Oct 2024   Prob (F-statistic):           9.61e-12
Time:                        18:38:09   Log-Likelihood:                -72.184
No. Observations:                  57   AIC:                             148.4
Df Residuals:                      55   BIC:                             152.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.2958      0.471     19.734      0.000       8.352      10.240
UI            -0.0745      0.009     -8.591      0.000      -0.092      -0.057
==============================================================================
Omnibus:                        3.407   Durbin-Watson:                   1.419
Prob(Omnibus):                  0.182   Jarque-Bera (JB):                2.436
Skew:                           0.428   Prob(JB):                        0.296
Kurtosis:                       3.543   Cond. No.                         221.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

MODEL 3 – FULL REGRESSION RESULTS

```
Model 3: Using ACO and UI
                            OLS Regression Results
==============================================================================
Dep. Variable:                   TODU   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.684
Method:                 Least Squares   F-statistic:                     61.66
Date:                Fri, 04 Oct 2024   Prob (F-statistic):           1.14e-14
Time:                        18:38:09   Log-Likelihood:                -62.548
No. Observations:                  57   AIC:                             131.1
Df Residuals:                      54   BIC:                             137.2
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.4263      1.119      3.955      0.000       2.182       6.670
ACO            3.7253      0.799      4.661      0.000       2.123       5.328
UI            -0.0395      0.011     -3.742      0.000      -0.061      -0.018
==============================================================================
Omnibus:                        6.144   Durbin-Watson:                   1.542
Prob(Omnibus):                  0.046   Jarque-Bera (JB):                5.214
Skew:                           0.700   Prob(JB):                       0.0738
Kurtosis:                       3.483   Cond. No.                         746.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

MODEL 4 – FULL REGRESSION RESULTS

```
Model 4: Using ACO, UI and ACO*UI
                            OLS Regression Results
==============================================================================
Dep. Variable:                   TODU   R-squared:                       0.698
Model:                            OLS   Adj. R-squared:                  0.681
Method:                 Least Squares   F-statistic:                     40.83
Date:                Fri, 04 Oct 2024   Prob (F-statistic):           8.24e-14
Time:                        18:38:09   Log-Likelihood:                -62.312
No. Observations:                  57   AIC:                             132.6
Df Residuals:                      53   BIC:                             140.8
Df Model:                           3
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  5.3285      1.764      3.020      0.004       1.790       8.867
ACO                    2.5989      1.878      1.384      0.172      -1.167       6.365
UI                    -0.0584      0.030     -1.921      0.060      -0.119       0.003
ACO_UI_interaction     0.0245      0.037      0.664      0.510      -0.050       0.099
==============================================================================
Omnibus:                        6.414   Durbin-Watson:                   1.471
Prob(Omnibus):                  0.040   Jarque-Bera (JB):                5.528
Skew:                           0.727   Prob(JB):                       0.0630
Kurtosis:                       3.462   Cond. No.                     1.75e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.75e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

MODEL 5 – FULL REGRESSION RESULTS

```
Model 5: Using all the variables
                            OLS Regression Results
==============================================================================
Dep. Variable:                   TODU   R-squared:                       0.704
Model:                            OLS   Adj. R-squared:                  0.675
Method:                 Least Squares   F-statistic:                     24.28
Date:                Fri, 04 Oct 2024   Prob (F-statistic):           2.04e-12
Time:                        19:39:45   Log-Likelihood:                -61.720
No. Observations:                  57   AIC:                             135.4
Df Residuals:                      51   BIC:                             147.7
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.8174      2.208      1.276      0.208      -1.616       7.251
ACO            3.6467      0.957      3.813      0.000       1.726       5.567
AHS            0.3237      0.412      0.785      0.436      -0.504       1.151
MI             0.0053      0.009      0.574      0.569      -0.013       0.024
SRI            0.0081      0.009      0.924      0.360      -0.010       0.026
UI            -0.0363      0.013     -2.720      0.009      -0.063      -0.010
==============================================================================
Omnibus:                        4.558   Durbin-Watson:                   1.598
Prob(Omnibus):                  0.102   Jarque-Bera (JB):                3.583
Skew:                           0.572   Prob(JB):                        0.167
Kurtosis:                       3.445   Cond. No.                     1.71e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```